УДК 004.9

АНАЛИЗ ЛЕКСИЧЕСКИХ ПАР ДЛЯ ГЕНЕРАЦИИ ДИАЛОГИЧЕСКОЙ РЕЧИ

Прозоров Н.Р. , Личаргин Д.В., Веретенникова А.В.
Научный руководитель – канд. тех. наук. Д.В. Личаргин, А.В. Веретенникова
Сибирский Федеральный Университет

В работе рассматривается проблема формирования корректного и осмысленного текста посредством использования программных систем.

На сегодняшний день широко распространены и разрабатываются разнообразные системы формирования высказываний различными программными системами: экспертными системами, программами электронного перевода, «ботами» (системами диалога с пользователем), синонимизаторами, программами генерации текстов по тематике «прогноз погоды», «технический справочник» и т.п.

Проблема является актуальной в связи с важностью развития принципов и систем искусственного интеллекта и потребностью формирования осмысленного текста с помощью средств вычислительной техники для различных практических приложений.

Проблема решается на стыке таких наук, как информатика, математика, системный анализ, лингвистика, философия, психология и пр.

Проблема исследуется со времен появления вычислительной техники и широко исследуется различными авторами, в частности Э. Кодда, А. Хомского, А.С.Нариньяни, Т. Винограда, М.В. Никитина, К. Шеннона, А.И. Пиотровского и многих других.

Однако вопрос требует дополнительных исследований в рамках анализа структуры естественных языков.

Цель данной работы состоит в том, чтобы дать анализ лексических пар для генерации диалогической речи.

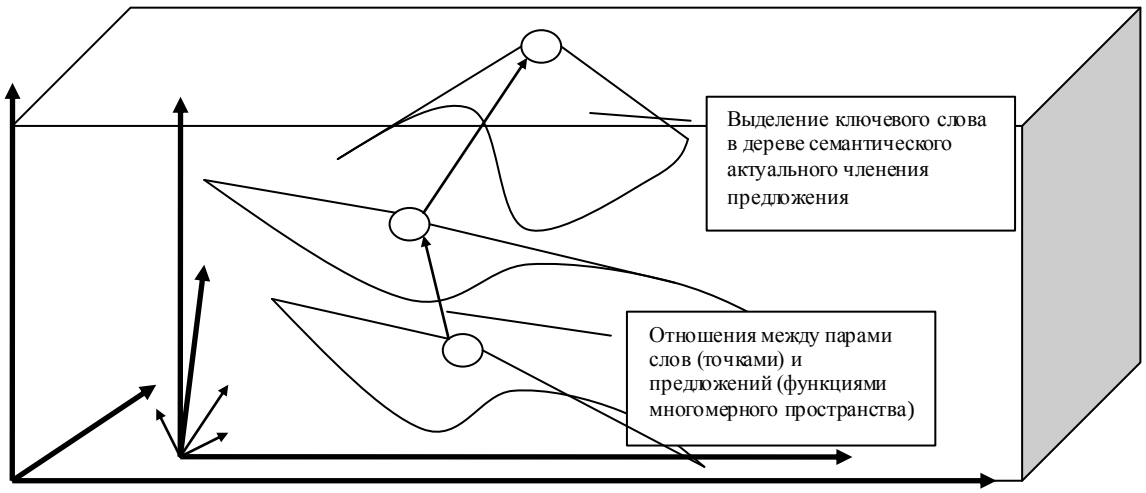

Рисунок 1. Модель лексико-грамматического пространства

Задачи данной работы заключаются:

1) в анализе классификации английского языка для ее последующего использования в качестве основы генерации осмысленного подмножества языка.
2) в анализе взаимосвязей между словами и выражениями в английском языке – их пар как векторов многомерного пространства слов языка и траекторий слов и предложений как цепочки или системы векторов.

Основная идея работы состоит в построении модели естественного языка на основе

многомерного представления слов и пар слов языка и в применении этой модели на примере английского языка с русским подстрочником.

Новизна данной работы подтверждается отсутствием полностью удовлетворительных и полных систем генерации осмысленной речи, не смотря на различные реализации решения этой проблемы в различных приближениях (программа Alice и др.)

Проблема формирования связного текста, в частности английского, является центральной задачей компьютерной лингвистики – дисциплины, лежащей на стыке информатики, математики, системного анализа, лингвистики, философии, психологии и пр. Решение задач семантики, дискретной математики, лингвистики и искусственного интеллекта направлены на прохождение теста Тьюринга с все более жесткими условиями, включающими в себя широкий набор слов, конструкций, фактов и эмуляции отношения к предмету разговора со стороны собеседника или выступающего.

Рассмотрим многомерное пространство объектов естественного языка: слов и выражений. Многие словосочетания могут быть сформированы правильно относительно грамматики, но не иметь семантического смысла. Допустим, фраза "See I" грамматически построена не верно, а фраза "I eat a hat" грамматически корректна, но не имеет семантического смысла, а фраза "I eat a pear" верна и в грамматическом, и в семантическом смысле.

Ниже приводится пример учета комбинаторики слов естественного языка, представленного в форме подстановочной таблицы, способной генерировать осмысленные фразы на английском языке.

Таблица 1

Срез многомерного пространства в виде подстановочной таблицы, построенной по методу Палмера

Зэ ... ***этот ...***	...-(e)s *...-(и)с/з* **...**	...-ing *...-иН* **...**	the ... *Зэ ...* ***этот ...***
cracker *крАекэ* ***взломщик программного обеспечения***	finish *фИниш* ***заканчивать***	optimise *оптимАйз* ***оптимизировать***	software *сОфтВеэ* ***программное обеспечение***
user *йУ:зэ* ***пользователь***	give up *гИв Ап* ***бросить***	improve *импрУ:в* ***улучшать***	file *фАйл* ***файл***
private user *прАйвит йУ:зэ* ***частный пользователь***	continue *кэнтИ:нуэ* ***продолжать***	maintain *мэнтЕйн* ***поддерживать в хорошем состоянии***	project *прОджэкт* ***проект***
client *клАйэнт* ***клиент***	control *кэнтрОл* ***контролировать***	make an error in *мЕйк эн Ероу ин* ***допускать ошибку в***	program *прОгрэм* ***программа***

Возможно построение многомерной базы данных со следующими координатами вектора понятийного описания:

v_1 = Части речи {«Артикль», «Прилагательное», «Существительное», «Глагол», ...};

v_2 = Члены предложения {«Определитель», «Определение», «Подлежащее», «Сказуемое», …};

$v_{3,3,1}$ = Лица {«1-ое», «2-ое», «3-ее», «Не определено»};

$v_{3,3,2}$ = Аспект {«Неопределенный», «Продолженный», «Совершенный», «Совершенный продолженный», «Не определен»};

$v_{3,1,1}$, $v_{3,1,2}$, … – Другие размерности, выраженные грамматическими категориями.

Далее, определим лексическое пространство языка (лексический куб) со следующими координатами:

l_1 = Порядок слов {Исполнитель, Действие, Реципиент, Получатель, Место, Время, Инструмент, Метод}

l_2 = Тема {Еда, одежда, тело, здание, группа людей, транспорт, ...}

1_3 = Варианты замены слов в предложение {to cook, to boil, to roast, to fry, to bake, …, to eat, to chew, …}

Все грамматические конструкции располагаются в ячейках многомерного массива данных – многомерного пространства слов языка. Координаты вектора, такие как, на пример, V[Глагол / Признак / Совершенный, ...], определяют ячейку с грамматической конструкцией "having + ГЛАГОЛ + -(e)d". Вектор V[Прилагательное / Предикат / Первое лицо, Превосходная степень , длинное слово, ...] определяет конструкцию "am the most + ПРИЛАГАТЕЛЬНОЕ". Реляционные таблицы, как часть этого многомерного массива, представлены в лингвистике в форме традиционных грамматических парадигм.

Таблица 2

Возможные отношения между словами со стороны шестимерного лексико-грамматического пространства

Название лексического и грамматического отношения	**Вектор многомерного пространства для слова 1**	**Вектор многомерного пространства для слова 2**	**Пример отношения**
Различие в частях речи	v[«Verb», B, C] + l[D, E, F]	v [«Noun», B, C] + l[D, E, F]	Love – to love
Различие в грамматической категории	v [A, B, «Singular»] + l[D, E, F]	v [A, B, «Plural»] + l[D, E, F]	Fan's – fans'
Различие в теме	v [A, B, C] + l[D, «Food, F= «Make»]	v [A, B, C] + l[D, «Clothes», F= «Make»]	Cook – sew
Различие в объекте	v[A, B, C] + l[D, E, F]	v [A, B, C] + l[D, E, F]	Start > launch
Антонимы	v [A, B, C] + l[D, E, F.G.H(disjunction level)]	v [A, B, C] + l[D, E, F.G.I (disjunction level)]	To be born – to live – to die – to revive
Гиперонимы	v [A, B, C] + l[D, E, F.G.H]	v [A, B, C] + l[D, E, F.G]	Mother – Parent
Гипонимы	v [A, B, C] + l[D, E, F.G]	v[A, B, C] + l[D, E, F.G.H]	Parent – Mother

Рассмотрим принцип сведения переходов между предложениями к переходам между словами на основе парсинга в форме дерева актуального членения предложения.

Традиционно актуальное членение предложений включает в себя деление на тему и рему, рема является ключевым словом в предложении, а тема относится ко всему тексту или его фрагменту. Таким образом, на вершине дерева актуального членения

предложения имеет место ключевое слово; на втором уровне дерева парсинга имеет место тема и рема; на третьем имеет место четверка: тема, связка, рема, модальность; на четвертом уровне добавляются обстоятельства, имеющие важную уточняющую функцию. На пятом уровне имеют место очевидные, понятные из контекста обстоятельства и конкретизация; на шестом – полупустые слова, уточняющие аспекты слов, указанных выше в дереве разбора.

Например,

0. Тема повествования: «суп»;
1. Ключевое слово: «вкуснятина» = «вкусный»;
2. Тема-Рема: «суп – вкуснятина» = «суп – вкусный»;
3. Тема-Рема-Связка-Модальность: «суп-вкусным-вышел-классно (очень хорошо)»;
4. Важная конкретизация: «…вкусным и профессиональным»;
5. Контекстуальная конкретизация: «суп, который готовила Аня, …»;
6. Аспекты понятий: «впечатление от супа, …, это просто восторг от вкусняшки, профессиональной штуки…»;
7. Различные эквивалентные преобразования, например, двойное отрицание.

Таким образом, одну и ту же мысль, что суп вкусный можно выразить астрономическим количеством более частных по смыслу и по форме фраз.

Приведем дополнительный пример: генерации дерева синонимичных по контексту фраз. Например,

0. Тема повествования: «автомобиль»;
1. Ключевое слово: «надежность»;
2. Тема-Рема: «автомобиль – надёжность» = «автомобиль – надёжный»;
3. Тема-Рема-Связка-Модальность: «автомобиль-надежным-сконструировали-профессионалы (хорошо)»;
4. Важная конкретизация: «…надежным и функциональным»;
5. Контекстуальная конкретизация: «автомобиль, который купил Пётр, …»;
6. Аспекты понятий: «оценка автомобиля, …, это является идеалом надёжности, комфортабельного дизайна…»;
7. Различные эквивалентные преобразования, например, двойное отрицание: «…нисколько не опасен», «нельзя не заметить….».

Приведем дополнительные примеры: генерации последовательностей фраз на английском языке.

1. Тема: Еда → Овощи Контекст: Петя готовит овощи в духовке со специями по книге рецептов (Первое предложение).
2. Вкусно→ Овощи – вкусные → Присутствие свежих овощей завораживает отличным вкусом (Второе предложение).
3. Пять часов → Овощи – в пять часов → Овощи исчезли с тарелок в пять часов → Ерунда, что овощи не исчезли с тарелок в пять часов (Третье предложение).
4. Кухня → Кухня – Еда →Кухня располагает к еде → Светлая (хорошая) кухня располагает к приятной (хорошей) еде (Четвертое предложение и т.д.).

Таким образом, от модели траекторий в виде цепочек пар слов естественного языка, как точек многомерного пространства, можно перейти к соответствующей траектории ключевых слов как вершин деревьев генерации каждого из вариантов синонимичных фраз языка.

Выводы. В заключении необходимо отметить, что проблема генерации логико-грамматических переходов между парами предложений нуждается в дальнейшем исследовании. Метод аналогии между переходами в виде пар слов и переходам между предложениями в виде дерева с одним ключевым словом на корне дерева актуального членения предложения является эффективным и нуждается в дальнейшем развитии.